University of Ljubljana, Faculty of Computer and Information Science

BERT, GPT, and T5 models

Prof Dr Marko Robnik-Šikonja

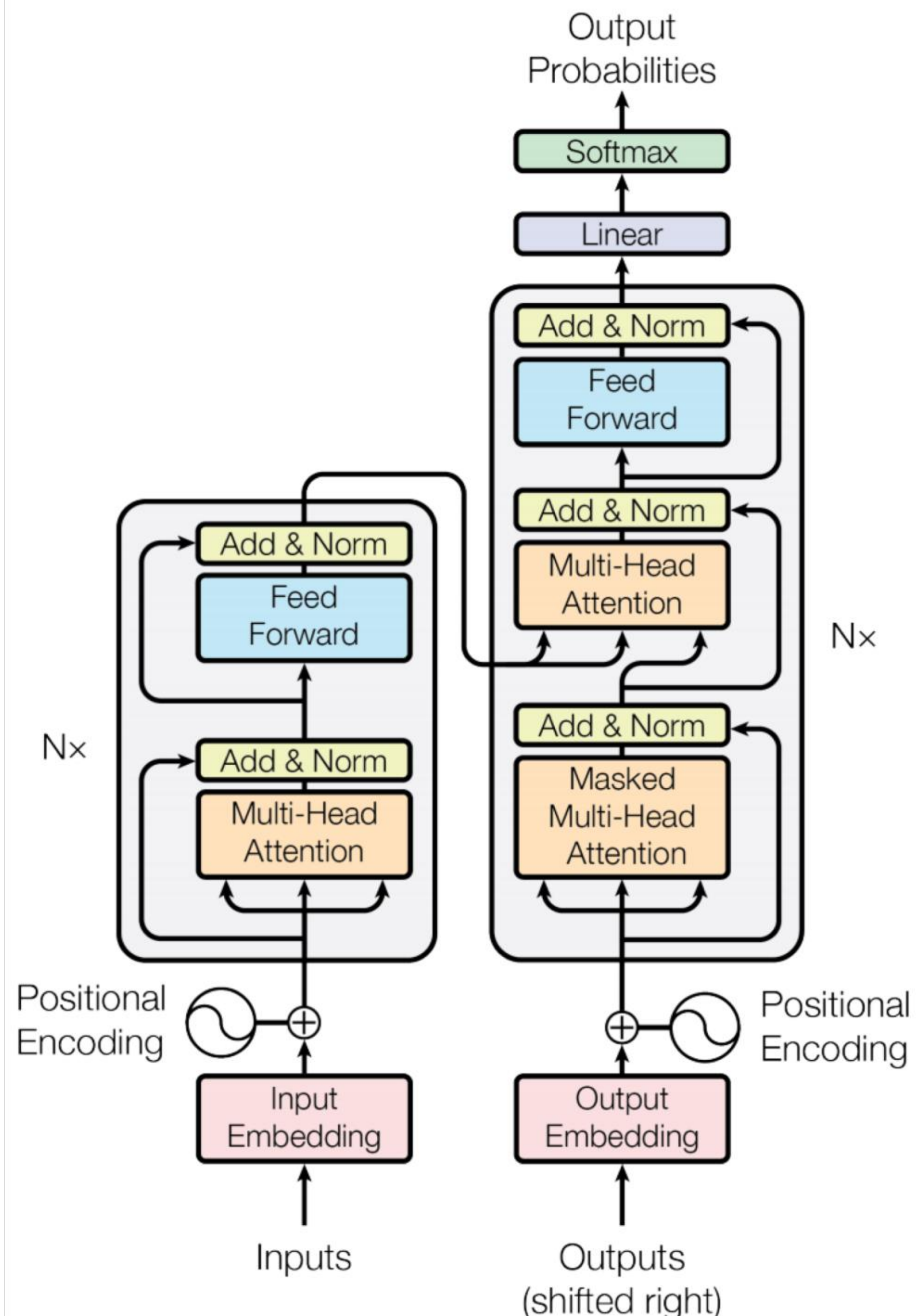
Natural Language Processing, Edition 2023

Contents

- BERT models
- GPT models
- T5 models
- other interesting transformers

- some slides by Jay Alammar, Jacob Devlin and Andrej Miščič

Transformer architecture revision



BERT

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- State-of-the-art pretained LM based on transformer architecture (only the encoder part)
- Idea:
  - use large unlabeled corpora and an auxiliary task to pretrain a model for a general language representation
  - fine-tune the model on a (possibly small) dataset for a specific downstream task
- presentation based on slides from Jacob Devlin and Jay Alammar

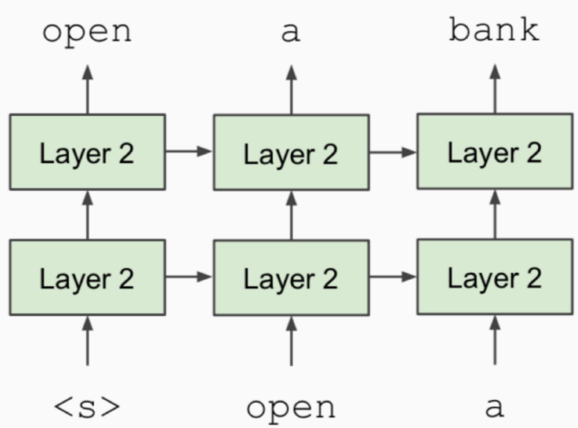
Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1, pp. 4171-4186.

BERT: motivation 1/3

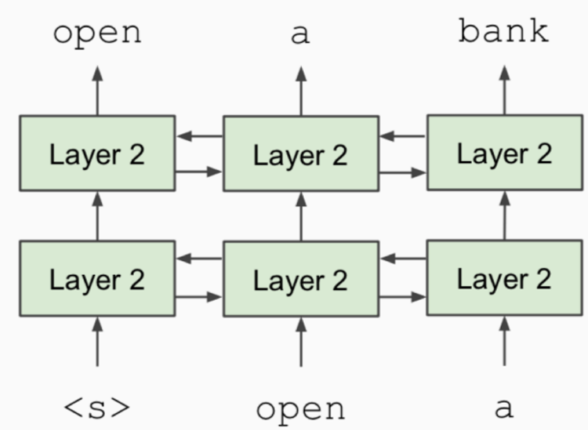
- **Problem**: Language models only use the left or right context, but language understanding is bidirectional.
- Why are LMs unidirectional?
  - Reason 1: Directionality is needed to generate a well-formed probability distribution.
    - We don’t care about this.
  - Reason 2: Words can “see themselves” in a bidirectional encoder.

BERT: motivation 2/3

Unidirectional context
Build representation incrementally



Bidirectional context
Words can “see themselves”



BERT: motivation 3/3

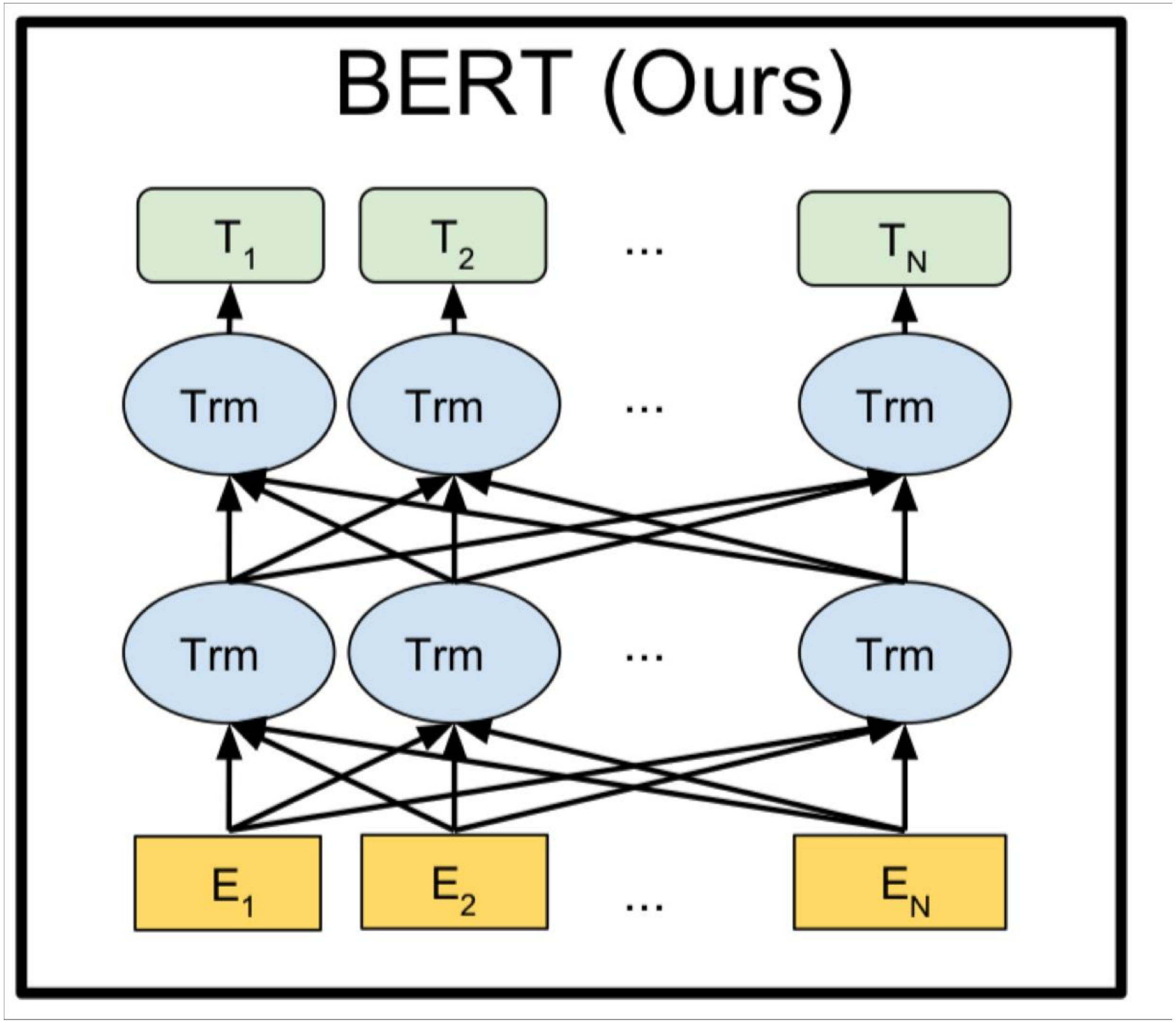
- **Solution**: Mask out *k*% of the input words, and then predict the masked words
- BERT uses *k* = 15%

store ↑ gallon ↑

the man went to the [MASK] to buy a [MASK] of milk

- Too little masking: Too expensive to train (not enough masks)
- Too much masking: Not enough context

BERT architecture



BERT uses several tasks

- besides masked LM, BERT learns relationships between sentences
- predict whether Sentence B is actual sentence that proceeds Sentence A, or a random sentence

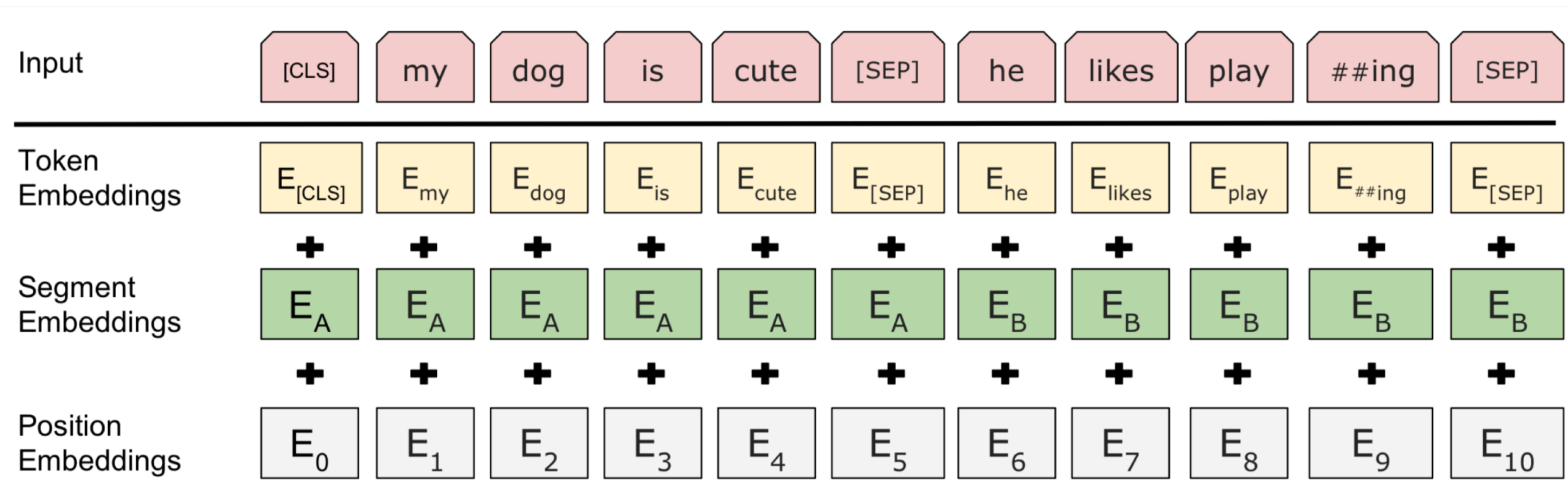
Sentence A = `The man went to the store.`
Sentence B = `He bought a gallon of milk.`
Label = `IsNextSentence`

Sentence A = `The man went to the store.`
Sentence B = `Penguins are flightless.`
Label = `NotNextSentence`

- some follow-up BERT-like models, e.g., RoBERTa, drop this task and claim better performance on downstream tasks

Sentence-pair encoding for BERT

- Token embeddings are word pieces (sub-word encoding)
- (Relatively) common words are in the vocabulary: *at, fairfax, 1910s*
- Other words are built from wordpieces: *hypatia = h ##yp ##ati ##a*
- Learned segmented embeddings represents each sentence
- Positional embedding is the same as for other transformer architectures

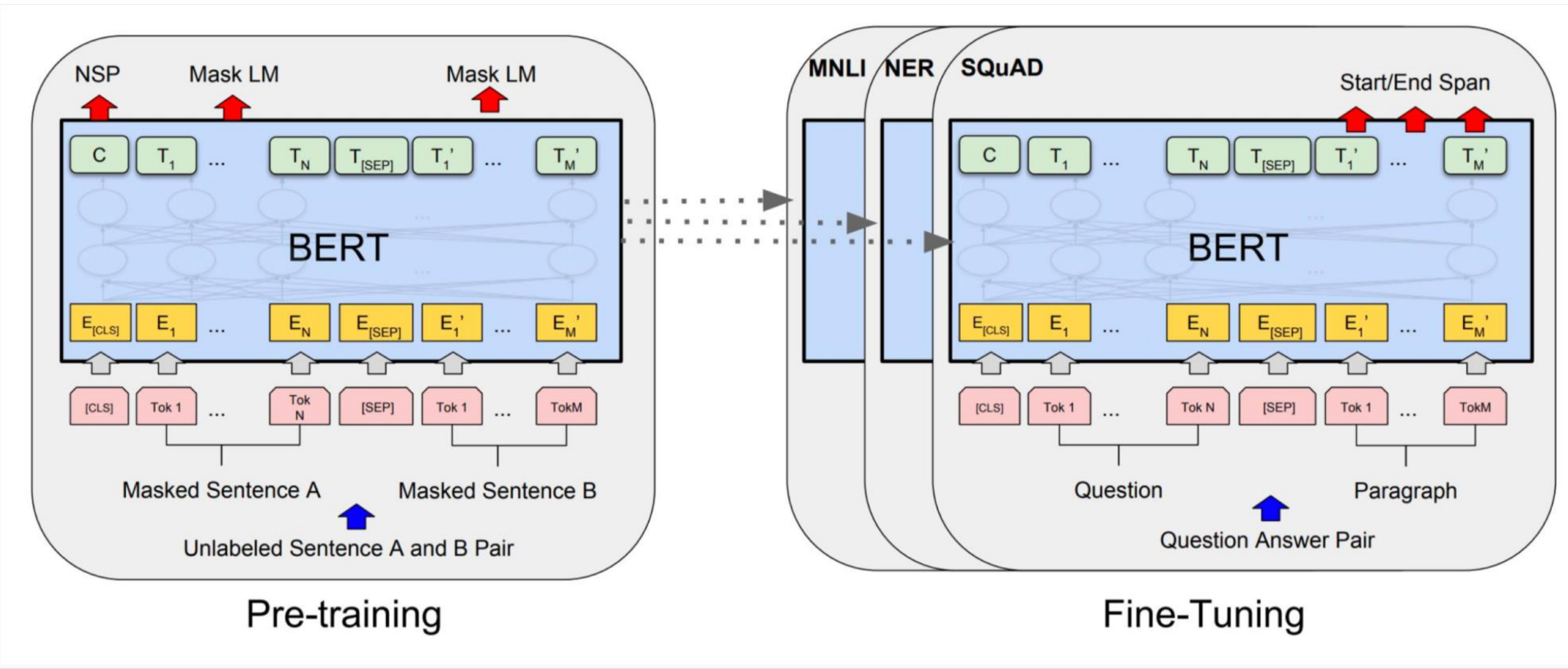


BERT training

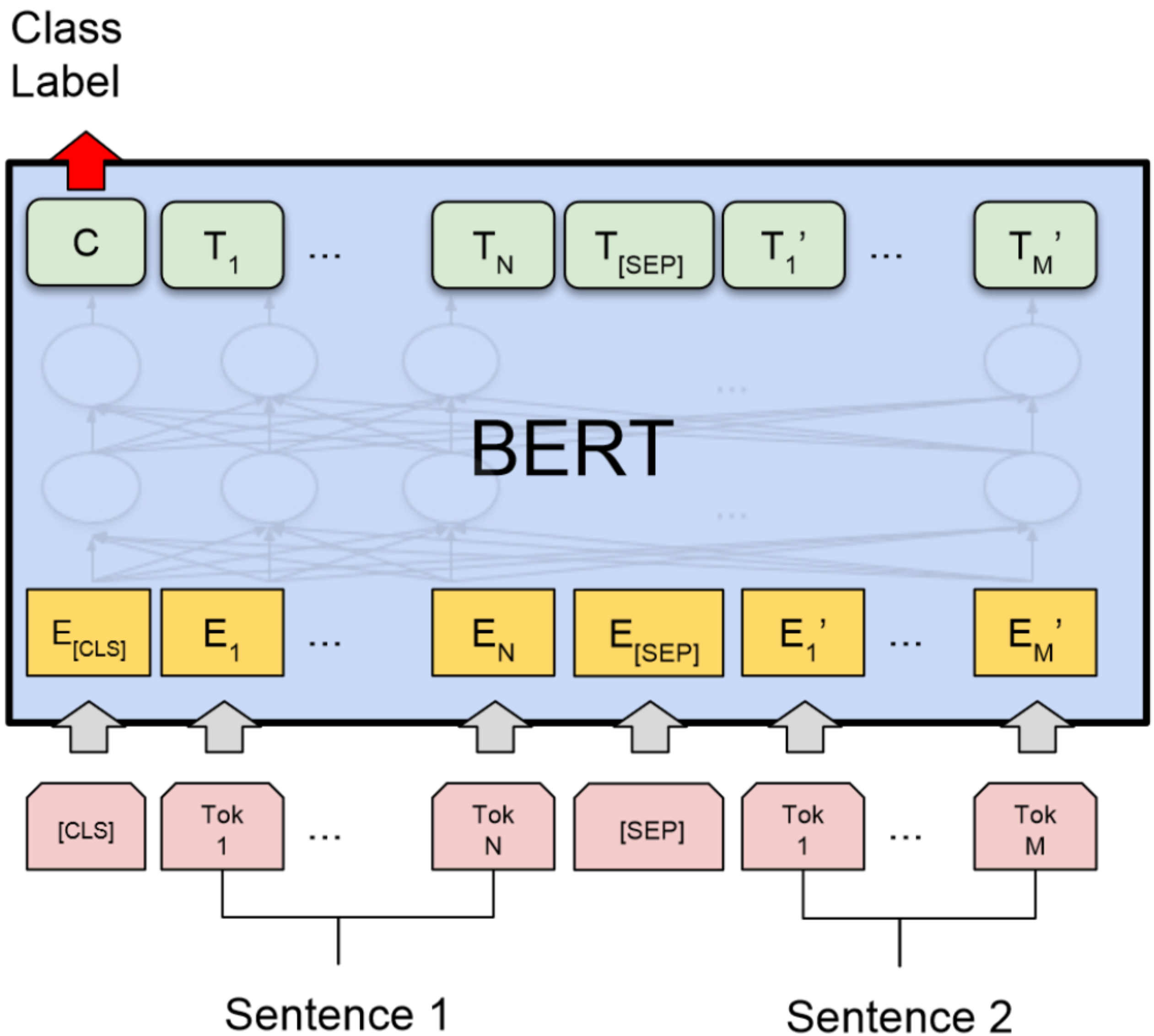
- Transformer encoder
- Self-attention ⇒ no locality bias
- Long-distance context has “equal opportunity”
- Single multiplication per layer ⇒ efficiency on GPU/TPU
- Trained on Wikipedia + BookCorpus
- English BERT was trained with 2 model sizes:
  - BERT-Base: 12-layer, 768-hidden parameters, 12-head, 110M parameters
  - BERT-Large: 24-layer, 1024-hidden parameters, 16-head, 340M parameters
- Trained on 4x4 or 8x8 TPU slice for 4 days

Use of BERT

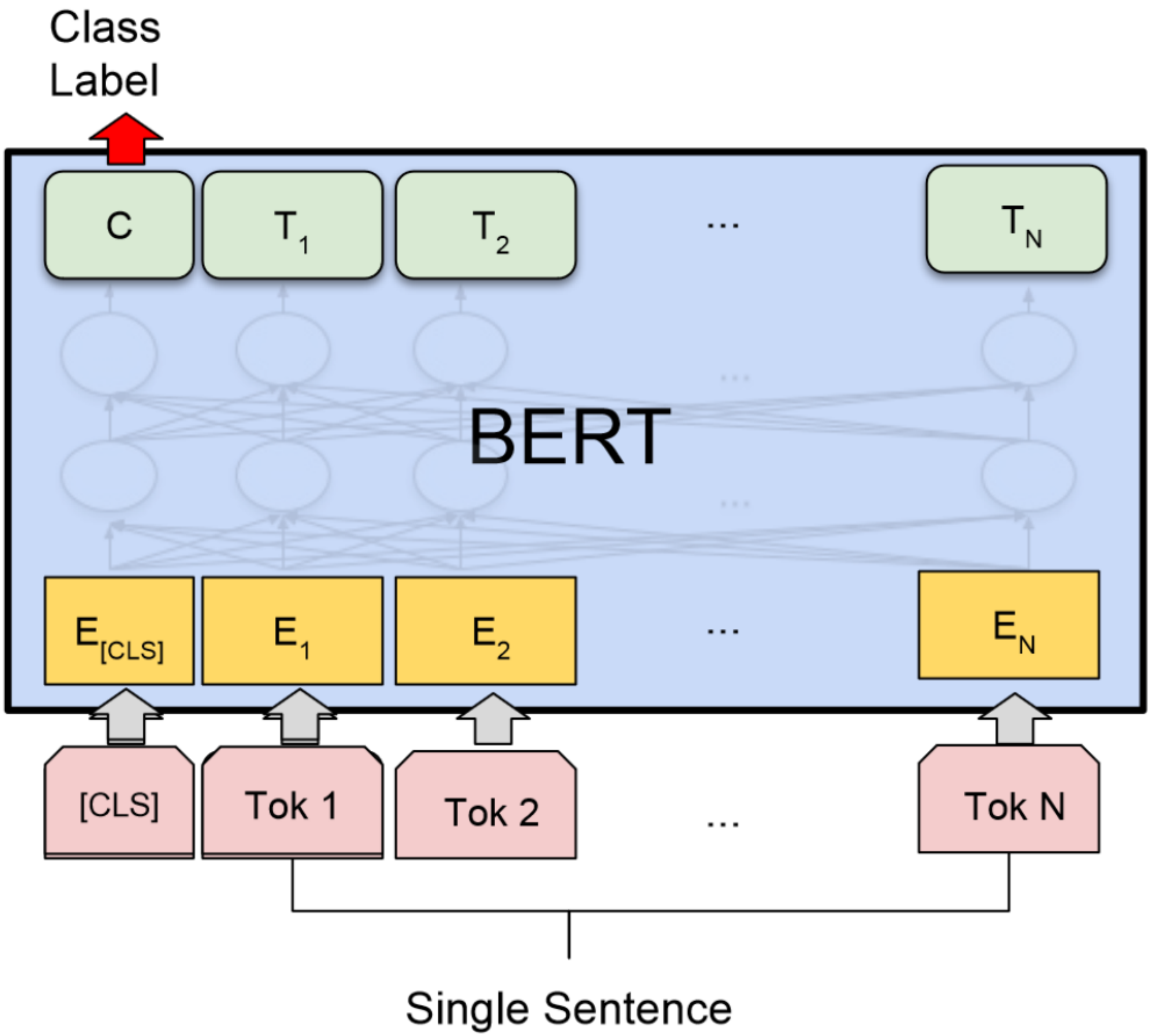
- train a classifier built on the top layer for each task that you fine-tune for, e.g., Q&A, NER, inference
- achieved state-of-the-art results for many tasks
- GLUE and SuperGLUE tasks for natural language understanding



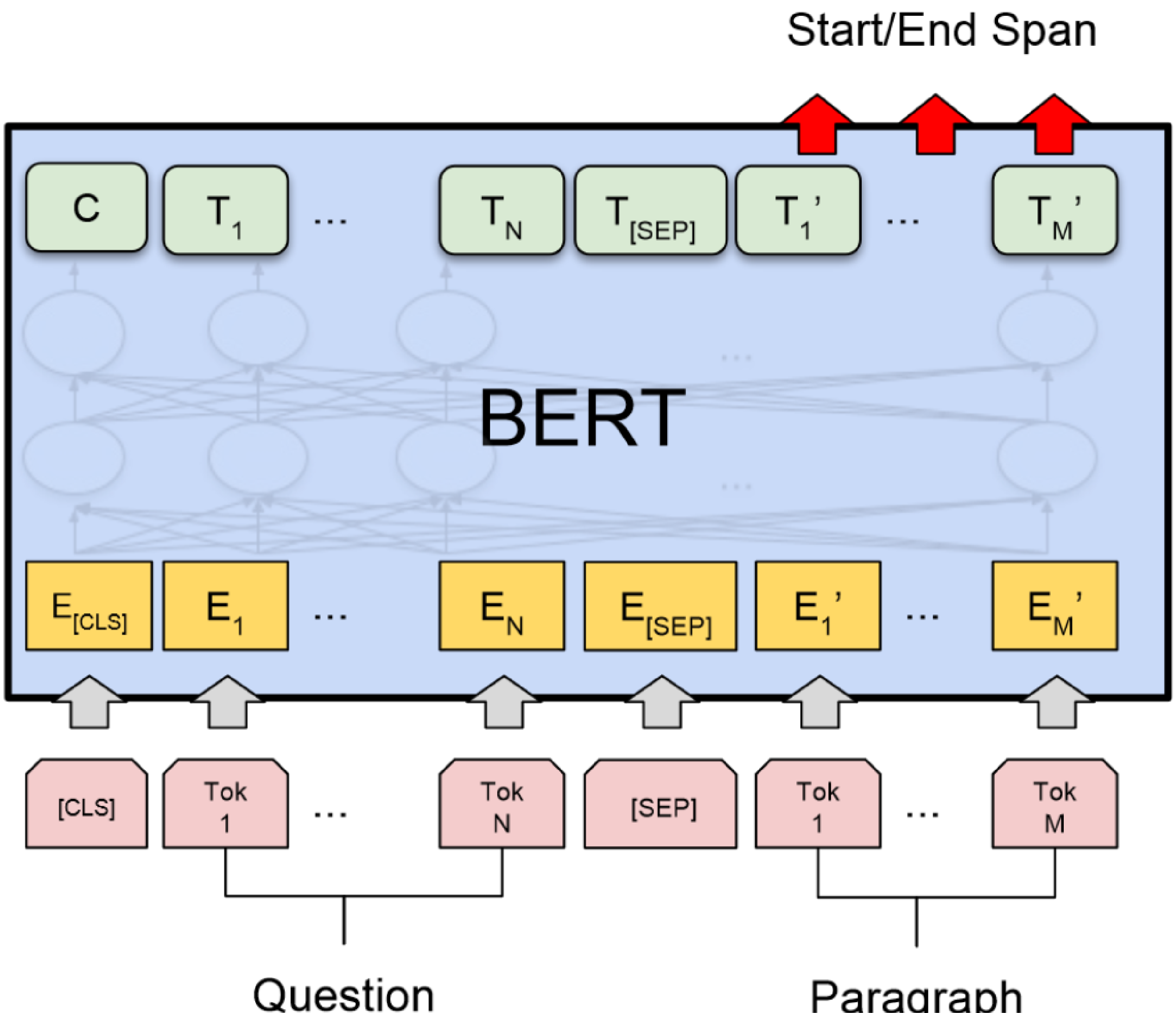
Two sentence classification using BERT-inference



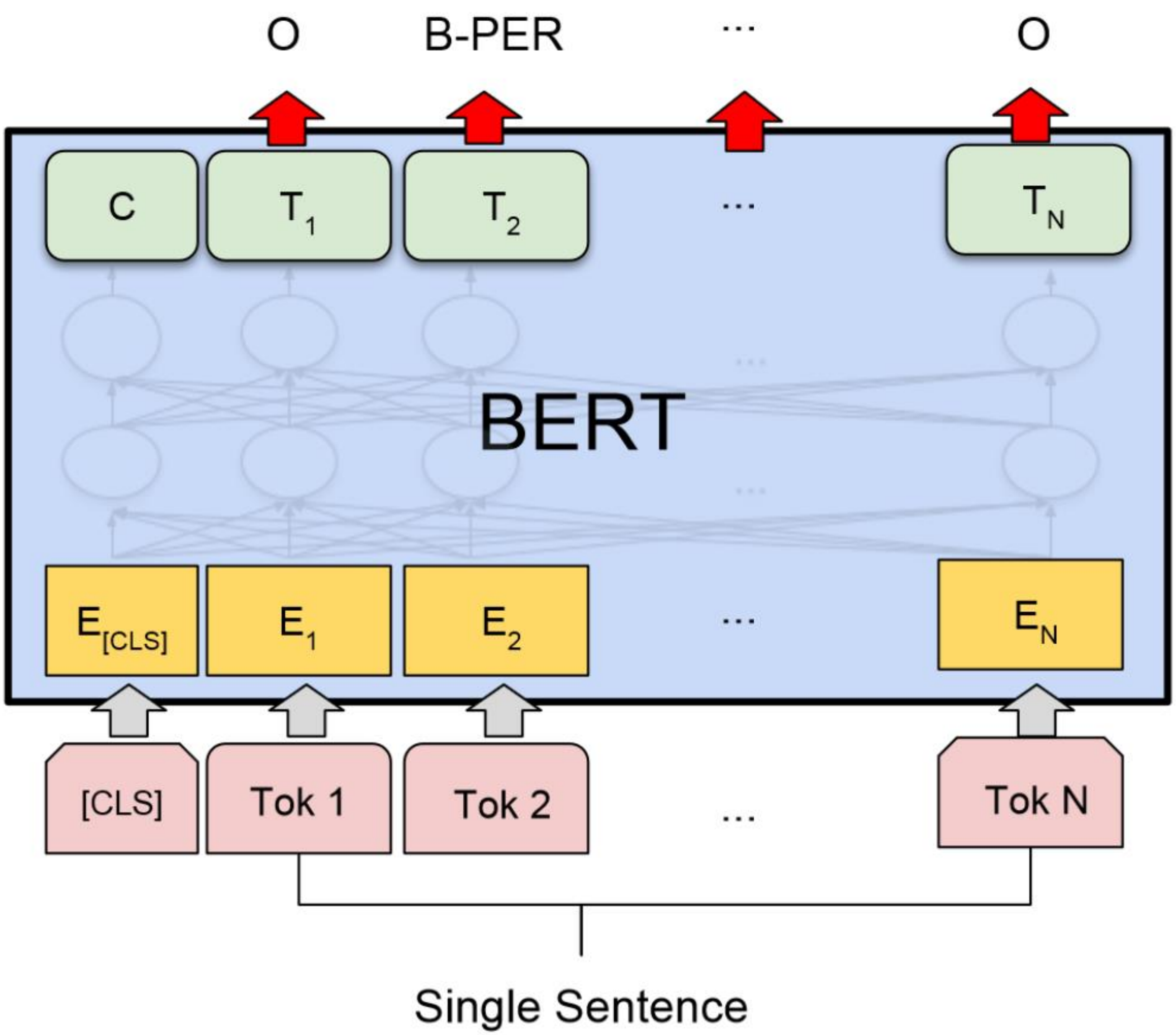
Sentence classification using BERT – sentiment, grammatical correctness



Questions and answers with BERT



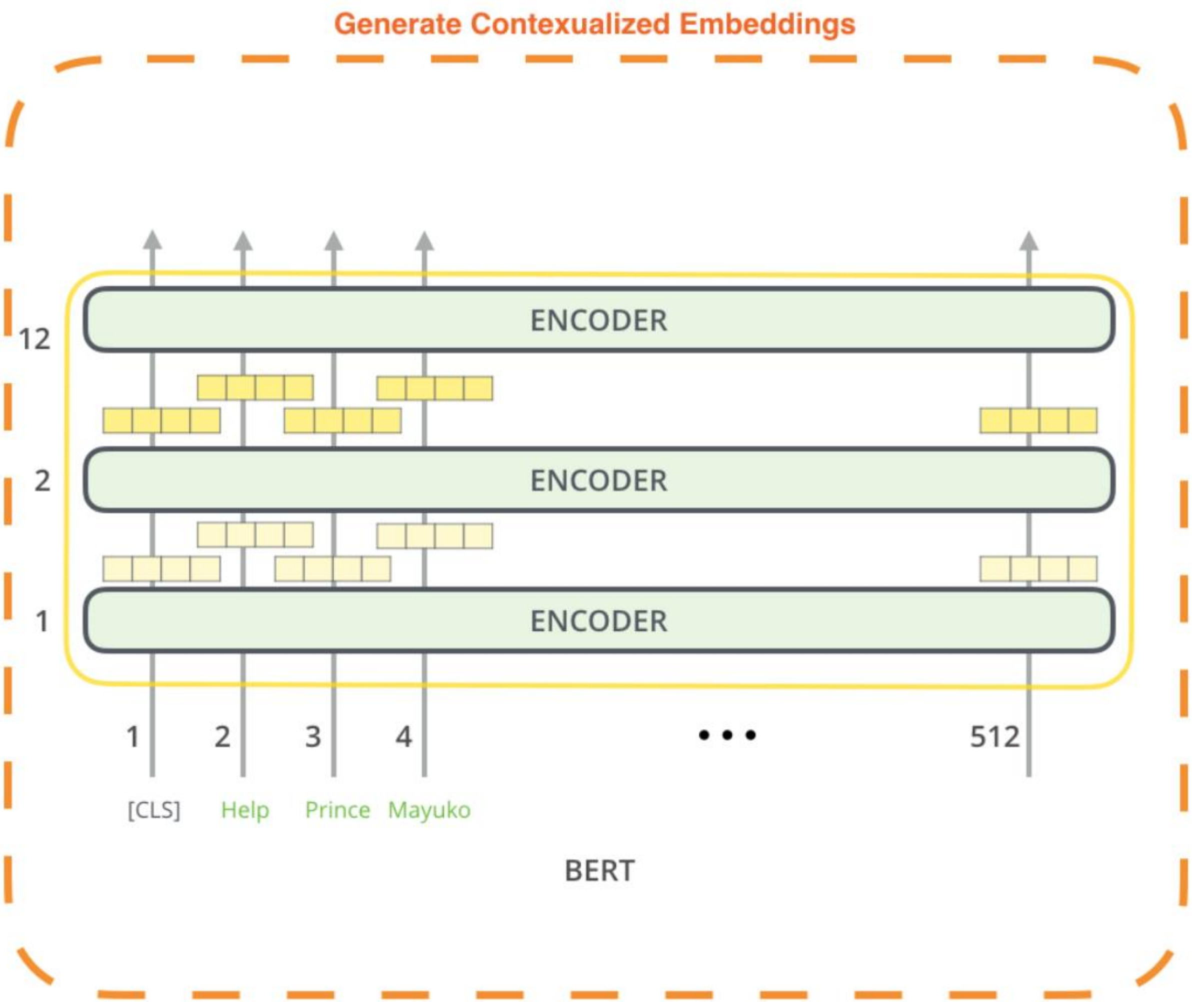
Sentence tagging with BERT- NER, POS tagging, SRL



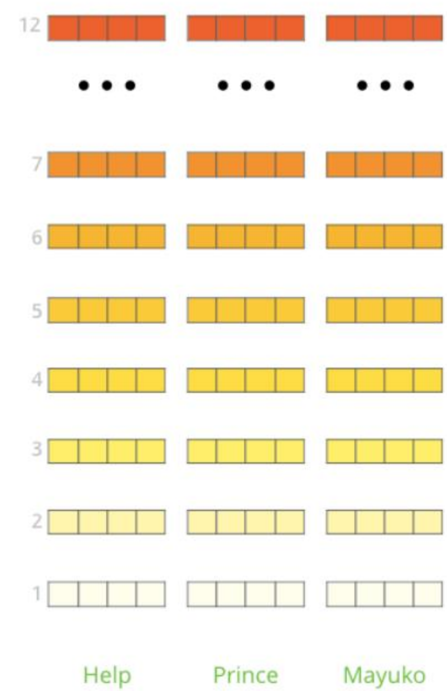
BERT can produce embeddings

- one can extract fixed size contextual vectors from BERT, achieving slightly lower accuracy than using the whole BERT as the first stage model

Layer-wise embeddings

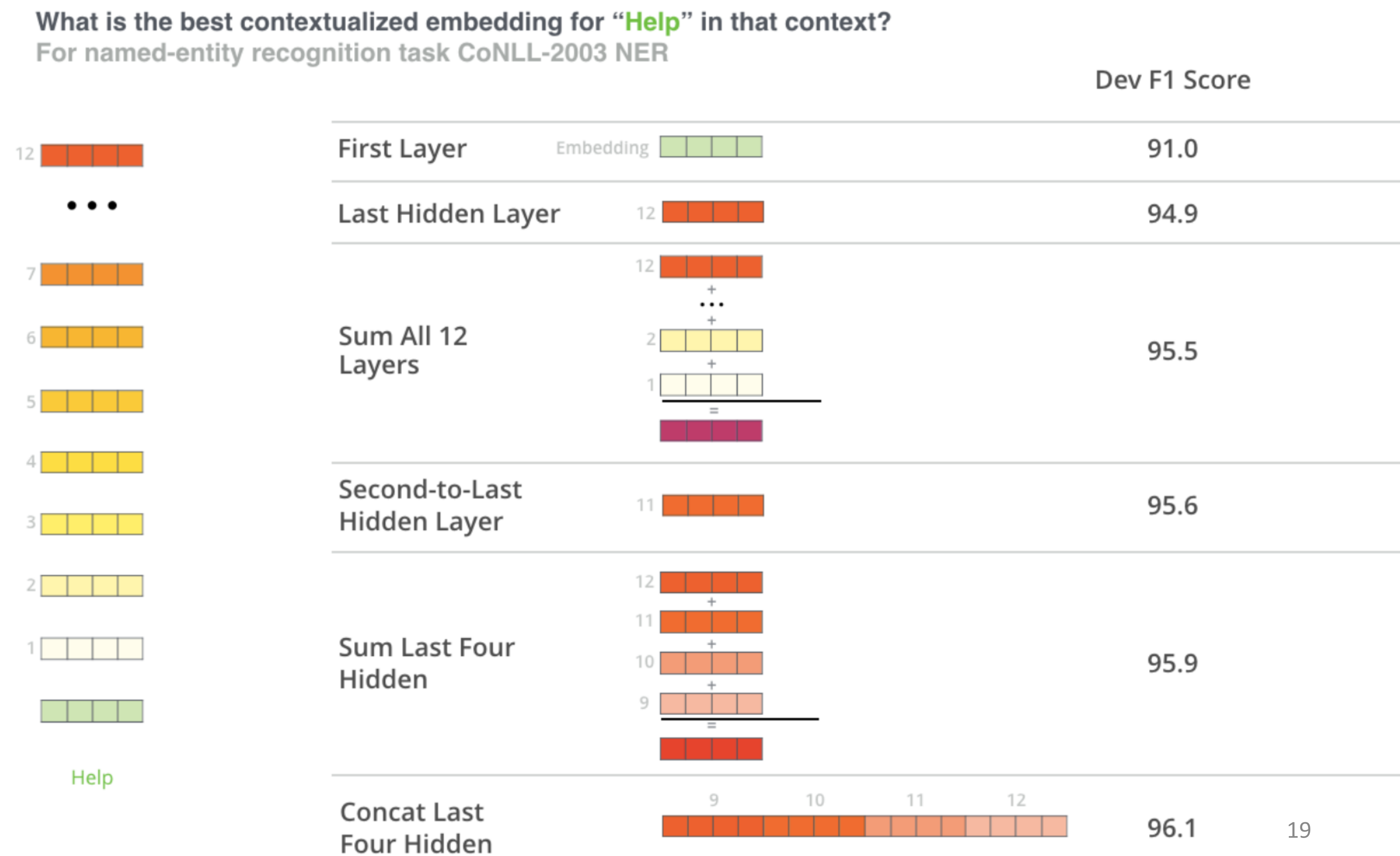


The output of each encoder layer along each token's path can be used as a feature representing that token.



But which one should we use?

Which layer of BERT to use as embeddings?



Examples of GLUE tasks

- GLUE benchmark is dominated by natural language inference tasks, but also has sentence similarity and sentiment

MultiNLI
Premise: Hills and mountains are especially sanctified in Jainism.
Hypothesis: Jainism hates nature.
Label: Contradiction

CoLA (Corpus of Linguistic Acceptability)
Sentence: The wagon rumbled down the road. Label: Acceptable
Sentence: The car honked down the road. Label: Unacceptable

SuperGLUE tasks

BoolQ - Boolean Questions
CB – Commitment Bank
COPA - Choice of Plausible Alternatives

MultiRC - Multi-Sentence Reading Comprehension
ReCoRD - Reading Comprehension with Commonsense Reasoning Dataset
RTE - Recognizing Textual Entailment
WiC - Word-in-Context
WSC - Winograd Schema Challeng

Table 2: Development set examples from the tasks in SuperGLUE. **Bold** text represents part of the example format for each task. Text in *italics* is part of the model input. <u>*Underlined*</u> text is specially marked in the input. Text in a `monospaced font` represents the expected model output.

BoolQ	**Passage:** *Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.* **Question:** *is barq's root beer a pepsi product* **Answer:** `No`
CB	**Text:** *B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?* **Hypothesis:** *they are setting a trend* **Entailment:** `Unknown`
COPA	**Premise:** *My body cast a shadow over the grass.* **Question:** *What's the CAUSE for this?* **Alternative 1:** *The sun was rising.* **Alternative 2:** *The grass was cut.* **Correct Alternative:** `1`

MultiRC

Paragraph: *Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week*
Question: *Did Susan's sick friend recover?* **Candidate answers:** *Yes, she recovered* (`T`), *No* (`F`), *Yes* (`T`), *No, she didn't recover* (`F`), *Yes, she was at Susan's party* (`T`)

ReCoRD

Paragraph: *(CNN) Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the State Electorcal Commission show. It was the fifth such vote on statehood. "Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood*
Query For one, they can truthfully say, "Don't blame me, I didn't vote for them, " when discussing the <placeholder> presidency **Correct Entities:** `US`

RTE

Text: *Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.*
Hypothesis: *Christopher Reeve had an accident.* **Entailment:** `False`

WiC

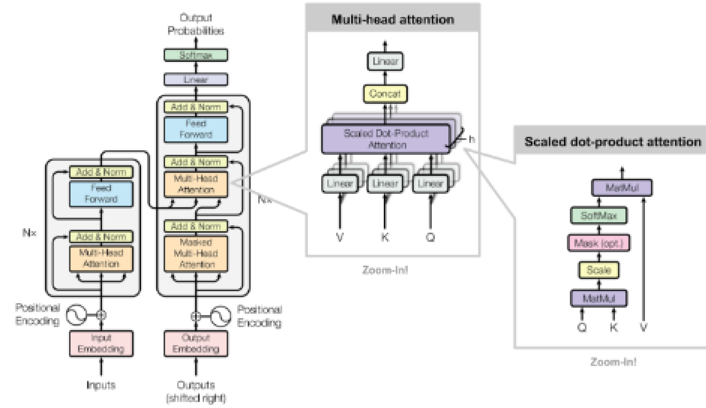
Context 1: *Room and board.* **Context 2:** *He nailed boards across the windows.*
Sense match: `False`

WSC

Text: *Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.* **Coreference:** `False`

Pretrained large language models (PLLMs)

- huge pretrained neural language models
- trained on large text corpora to capture relations in language
- finetuned to specific tasks
- publicly available BERT-like models
- for Slovene: fastText, ELMo, SloBERTa, CroSloEngual BERT, SlEng BERT, SloT5, Slo GPT
- for Croatian: fastText, ELMo, BERTić, CroSloEngual BERT
- on Clarin.si and HuggingFace
- hundreds of papers investigating BERT-like models in major ML & NLP conferences



- Ulčar, M., & Robnik-Šikonja, M. (2020). High Quality ELMo Embeddings for Seven Less-Resourced Languages. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 4731-4738).
- Ulčar, M. and Robnik-Šikonja, M., 2021. SloBERTa: Slovene monolingual large pretrained masked language model. *Proceedings of SI-KDD within the Information Society 2021*, pp.17-20.
- Ljubešić, N., & Lauc, D. (2021). BERTić-The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing (pp. 37-42).
- Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT. In *International Conference on Text, Speech, and Dialogue* (pp. 104-111).
- Ulčar, M. & Robnik-Šikonja, M. (2023) Sequence-to-sequence pretraining for a less-resourced Slovenian language.
- Frontiers in Artificial Intelligence, Section on Language and Computation, Volume 6 – 2023, https://doi.org/10.3389/frai.2023.932519

Multilingual PLMs

- Pretrained on multiple languages simultaneously
- multilingual BERT supports 104 languages by training on Wikipedia
- XLM-R was trained on 2.5 TB of texts
- these models allow cross-lingual transfer
- solve problem of insufficient training resources for less-resourced languages
- zero-shot transfer and few-shot transfer

SloBERTa

- Currently the best Slovene LM
- Between a few hundred and thousands downloads from HuggingFace
- Training set: 3.41 B words (corpora Gigafida, KAS, partially Janes, siParl, slWaC)
- Training duration: 4 weeks on Nvidia DGX A100 using 4xGPU
- An example of direct use:
  - <mask> je najlepše mesto na svetu.
  - Odgovori: Ljubljana, Barcelona, London, Madrid, To

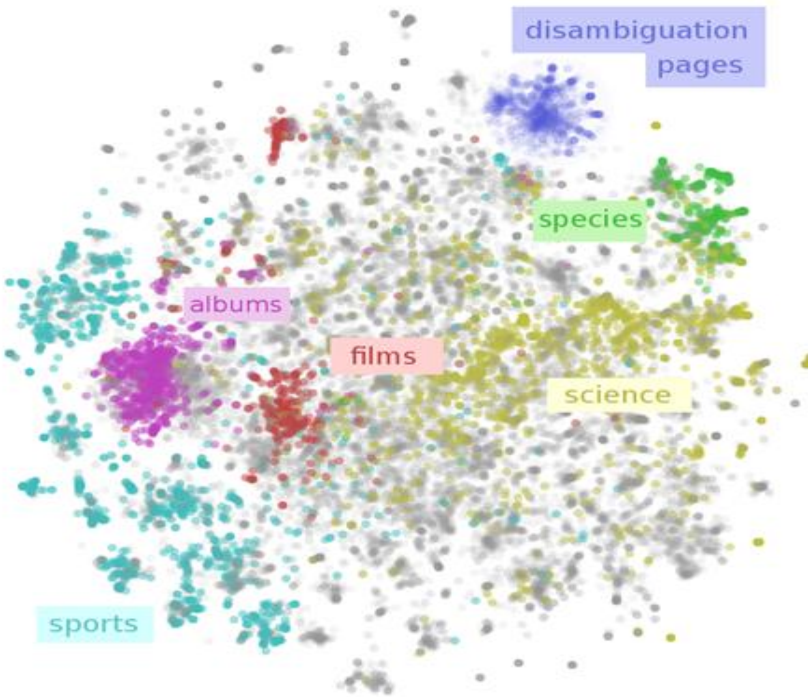
ULČAR, Matej, ROBNIK ŠIKONJA, Marko. (2021) SloBERTa: Slovene monolingual large pretrained masked language model. V IS *Data Mining and Data Warehouses - SiKDD* 2021, str. 17-20

Cross-lingual transfer

Explicit alignment of vector spaces

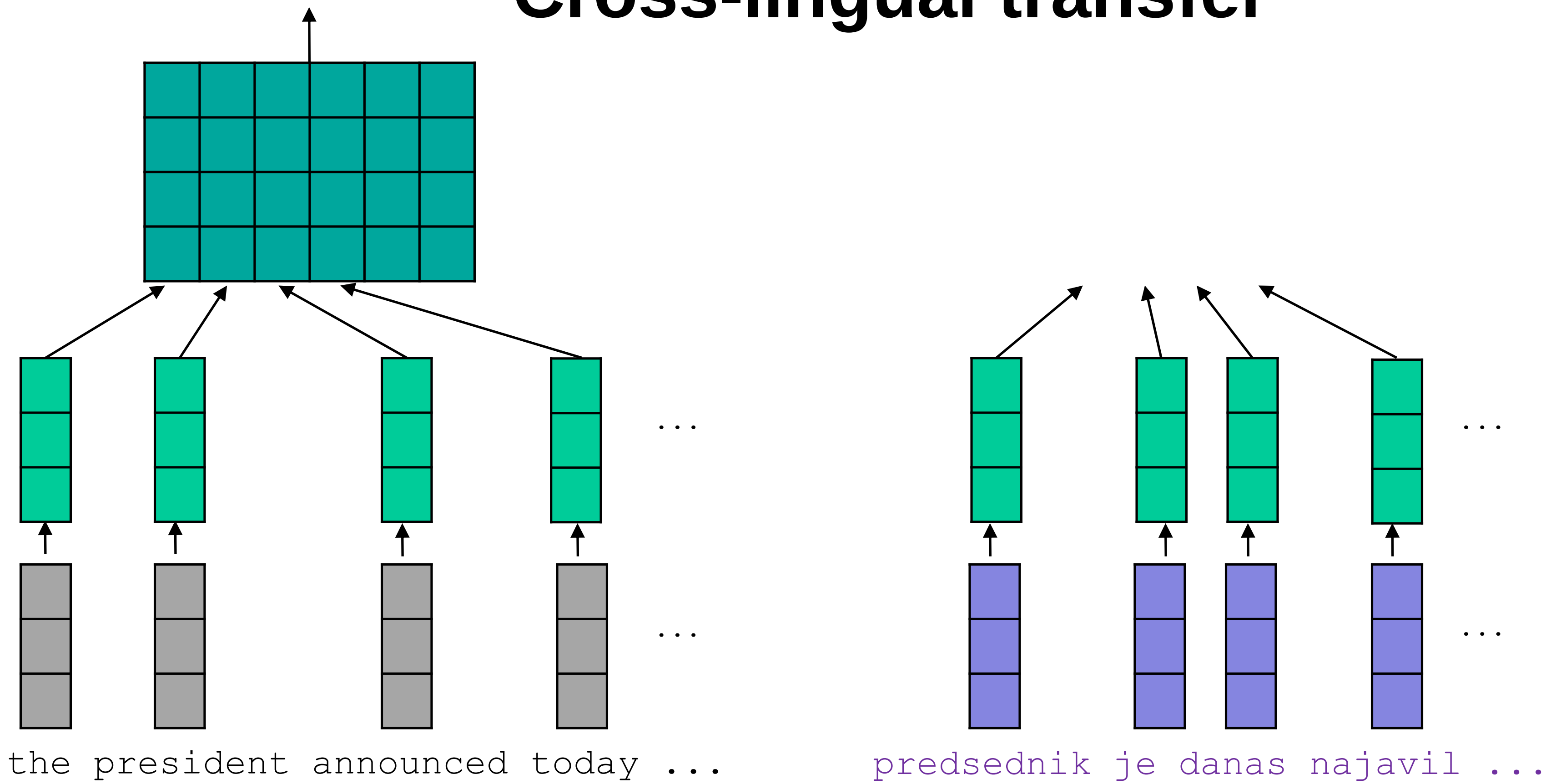
$WS \approx E$

Using multilinguak LLMs directly

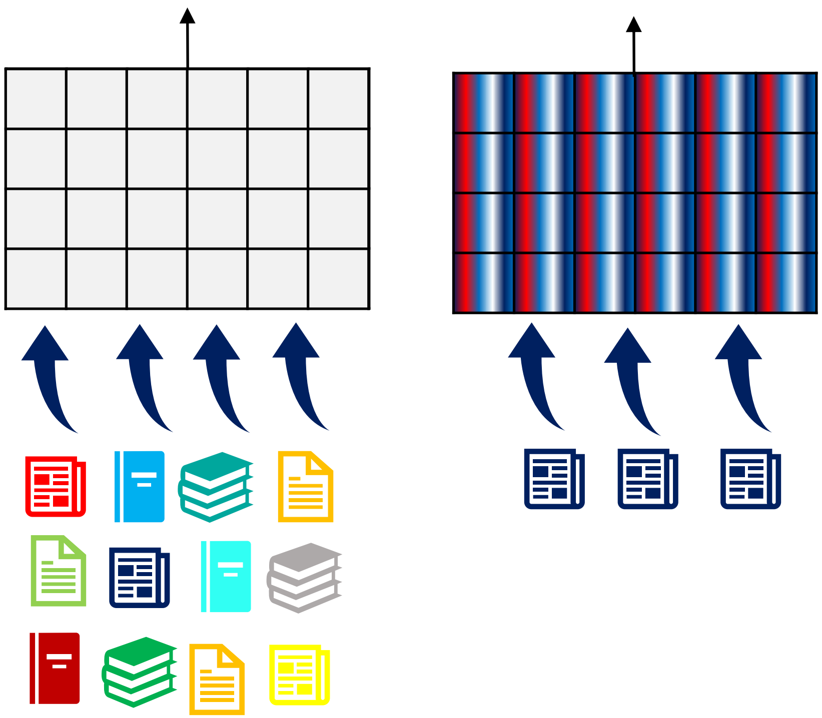


Ulčar, M. and Robnik-Šikonja, M., 2022. Cross-lingual alignments of ELMo contextual embeddings. *Neural Computing and Applications*, *34*(15), pp.13043-13061.

Cross-lingual transfer



Using multilingual models



predsednik je danas najavil ...

Pretraining   Fine-tuning   Classification

Zero-shot transfer and few-shot transfer

University of Ljubljana
Faculty of Computer and Information Science

Why not only multilingual?

- performance on many tasks drops with more languages
- results for a few tasks in Slovene (Named Entity Recognition – NER,Part-of-Speech Tagging – POS, Dependency Parsing – DP, Sentiment Analysis – SA, Word Analogy – WA)

Model	NER	POS	DP	SA	WA
mBERT	0.885	0.984	0.681	0.576	0.061
XLM-R	0.912	0.988	0.793	0.604	0.146
SloBERTa	**0.933**	**0.991**	**0.844**	**0.623**	**0.405**

Ulčar, M., Žagar, A., Armendariz, C. S., Repar, A., Pollak, S., Purver, M., & Robnik-Šikonja, M. (2021). Evaluation of contextual embeddings on less-resourced languages. *arXiv preprint arXiv:2107.10614*.

Dictionaries in PLMs

- Tokenization depends on the dictionary
- The dictionary is constructed statistically (SentencePiece algorithm)

- Sentence: “Letenje je bilo predmet precej starodavnih zgodb.”
- SloBERTa:

'▁Le', 'tenje', '▁je', '▁bilo', '▁predmet', '▁precej', '▁staroda', 'vnih', '▁zgodb', '.'

- mBERT:

'Let', '##en', '##je', 'je', 'bilo', 'pred', '##met', 'pre', '##cej', 'star', '##oda', '##vnih', 'z', '##go', '##d', '##b', '.'

Ulčar, M., & Robnik-Šikonja, M. (2021) Training dataset and dictionary sizes matter in BERT models: The case of Baltic languages. In *International Conference on Analysis of Images, Social Networks and Texts* (pp. 162-172)

Trade off: trilingual models

- BERT trained with only a few languages
- more data for training
- more specific dictionary
- good for cross-lingual transfer
- Trilingual models
  - CroSloEngual BERT
  - FinEst BERT
  - LitLat BERT
- SlavBERT (ru, pl, cs, bg; DeepPavlov)

Model	NER	POS	DP	SA	WA
mBERT	0.885	0.984	0.681	0.576	0.061
XLM-R	0.912	0.988	0.793	0.604	0.146
CSE-BERT	0.928	0.990	**0.854**	0.610	0.195
SloBERTa	**0.933**	**0.991**	0.844	**0.623**	**0.405**

- Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT. In International Conference on Text, Speech, and Dialogue (pp. 104-111).

University of Ljubljana
Faculty of Computer and Information Science

XL transfer in classsifictaion

- Excellent XL transfer between similar languages like Slovene and Croatian

sentiment analysis

		LASER		mBERT		CSE BERT		Both target	
Source	**Target**	$\bar{F}_1$	**CA**	$\bar{F}_1$	**CA**	$\bar{F}_1$	**CA**	$\bar{F}_1$	**CA**
Croatian	Slovene	0.53	0.53	0.53	0.54	**0.61**	**0.60**	0.60	0.60
Croatian	English	**0.63**	0.63	**0.63**	**0.66**	0.62	0.64	0.62	0.65
English	Slovene	0.54	0.57	0.50	0.53	**0.59**	**0.57**	0.60	0.60
English	Croatian	0.62	**0.67**	0.67	0.63	**0.73**	**0.67**	0.73	0.68
Slovene	English	0.63	0.64	**0.65**	**0.67**	0.63	0.64	0.62	0.65
Slovene	Croatian	0.70	0.65	0.64	0.63	**0.73**	**0.69**	0.73	0.68
Croatian English	Slovene	0.54	0.54	0.53	0.54	**0.60**	**0.58**	0.60	0.60
Croatian Slovene	English	0.62	0.61	**0.65**	**0.67**	0.63	0.65	0.62	0.65
English Slovene	Croatian	0.64	0.68	0.63	0.63	**0.68**	**0.70**	0.73	0.68
Average performance gap		0.04	0.03	0.04	0.03	0.00	0.01		

idiom detection

Language	Slovene ELMo	mBERT	Default F_1
Slovene	0.8163	0.8359	0.667
Croatian	0.9191	0.8970	0.667
Polish	0.2863	0.6987	0.667

- Robnik-Šikonja, M., Reba, K., & Mozetič, I. (2021). Cross-lingual transfer of sentiment classifiers. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, *9*(1), 1-25.
- Škvorc, T., Gantar, P., & Robnik-Šikonja, M. (2022). MICE: mining idioms with contextual embeddings. *Knowledge-Based Systems*, *235*, 107606.

What PLMs learn?

- We would like to travel to [MASK], ki je najlepši otok v Mediteranu.

SloBERTa: ..., Slovenija, I, Koper, Slovenia

CSE-BERT: Hvar, Rab, Cres, Malta, Brač

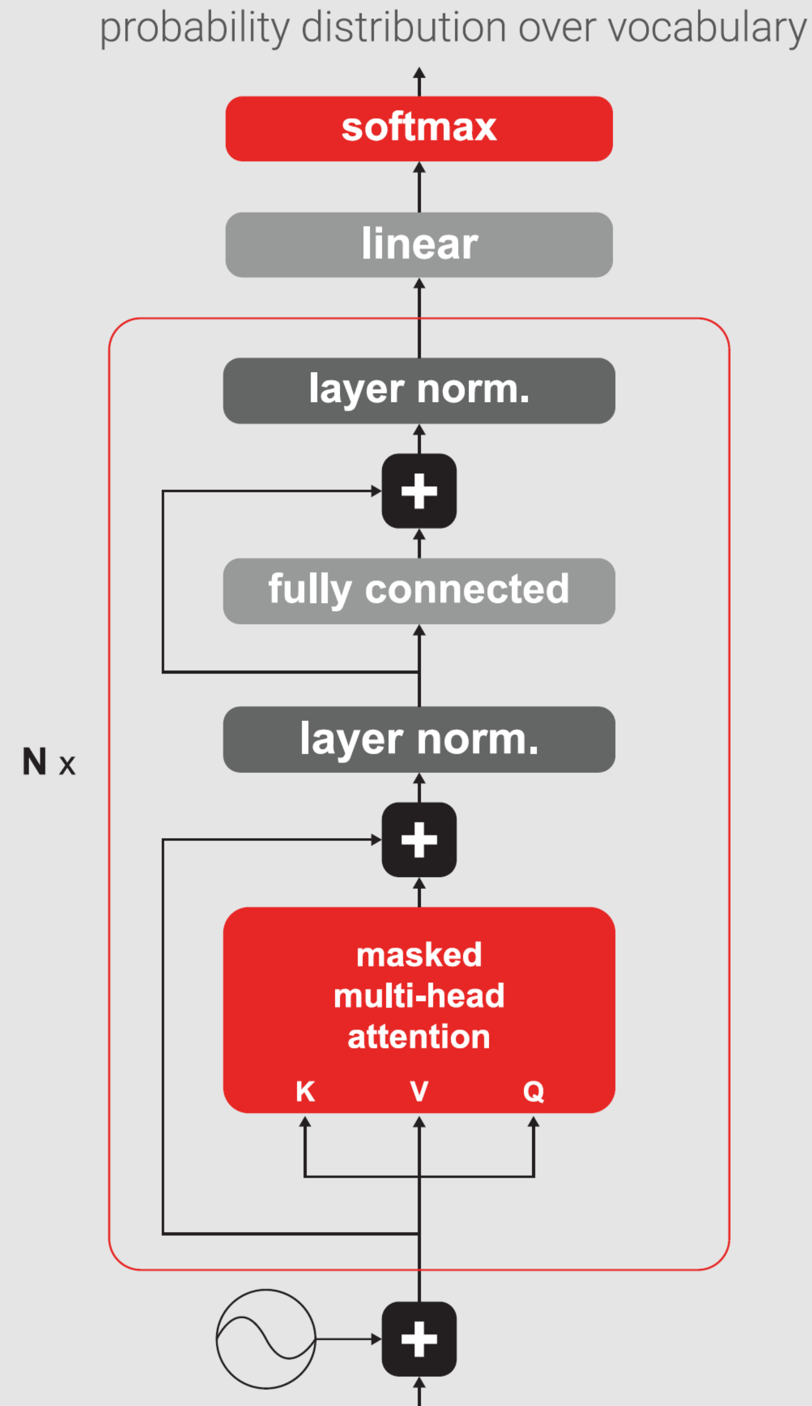
XLM-R: Mallorca, Tenerife, otok, Ibiza, Zadar

mBERT: Ibiza, Gibraltar, Tenerife, Mediterranean, Madeira

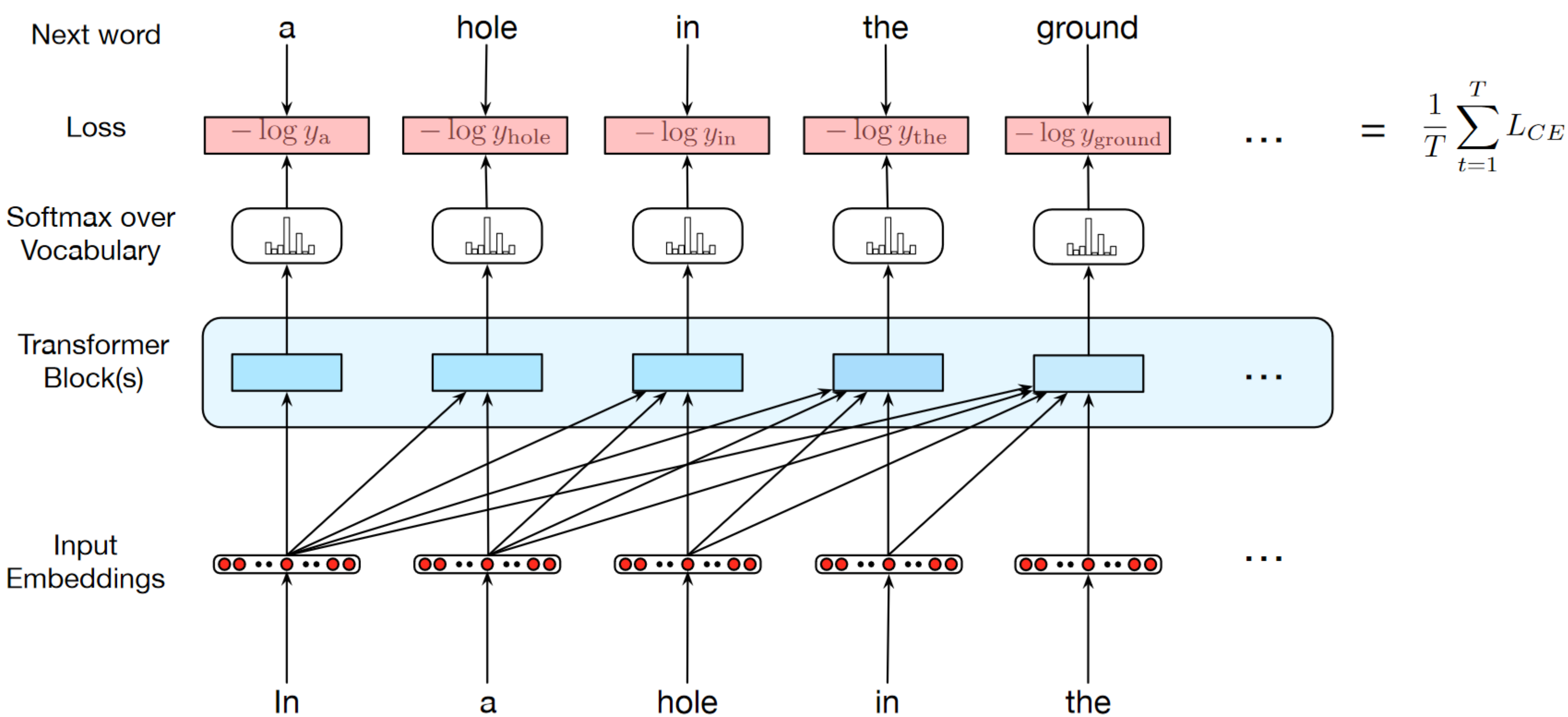
BERT (en): Belgrade, Italy, Serbia, Prague, Sarajevo

GPT family

- GPT: Generative Pre-trained Transformers
- use only the decoder part of transformer
- pretrained for language modeling (predicting the next word given the context)
- Shortcoming: unidirectional, does not incorporate bidirectionality
- “What are <u>those</u>?” he said while looking at my <u>crocs</u>.



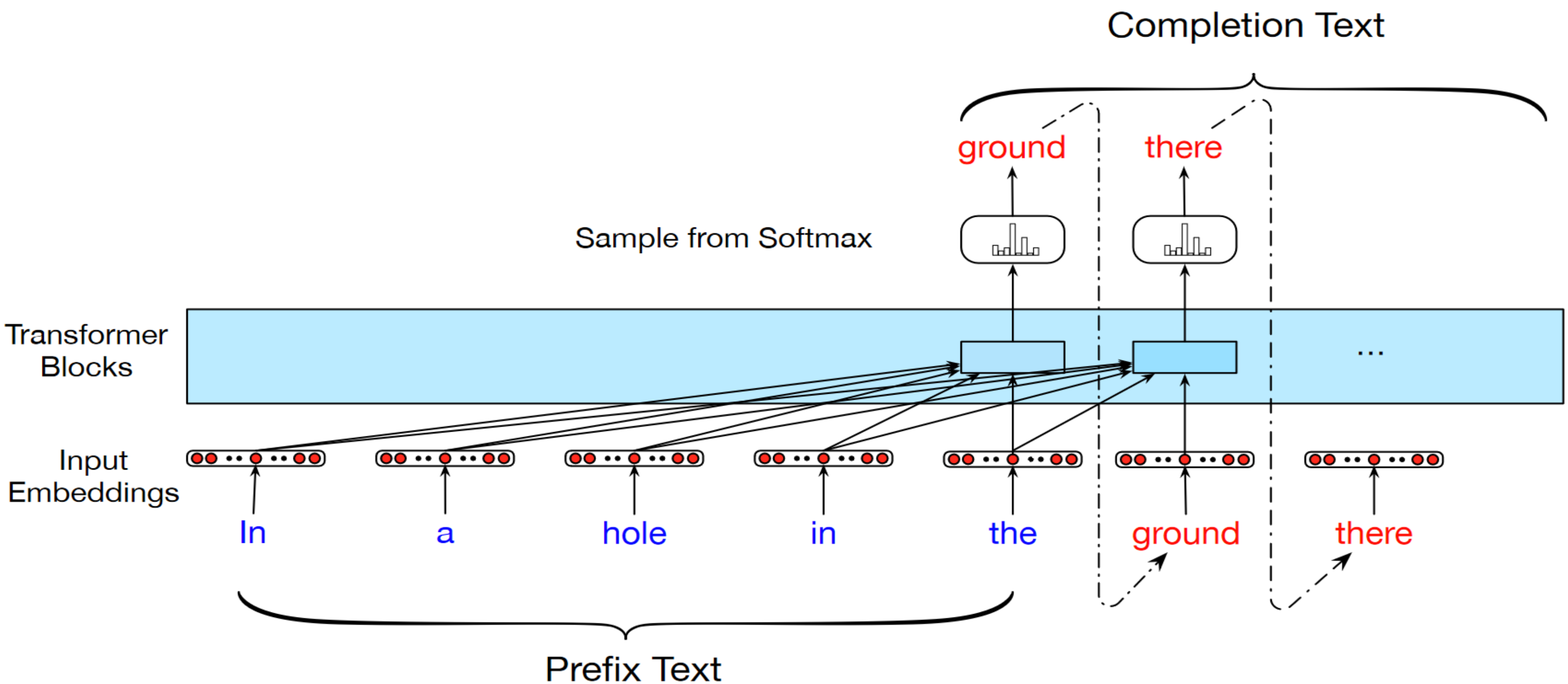
Transformer as a language model



- Can be computed in parallel

Autoregressive generators

- priming the generator with the context



- can be used also in summarization, QA and other generative tasks

GPT-2 and GPT-3

- few architectural changes, layer norm now applied to input of each subblock
- GPT-3 also uses some sparse attention layers
- more data, larger batch sizes (GPT-3 uses batch size of 3.2M)
- the models are scaled:

GPT-2:
48 layers, 25 heads
$d_m = 1600$, $d = 64$
context size = 1024
~ 1.5B parameters

GPT-3:
96 layers, 96 heads
$d_m = 12288$, $d = 128$
context size = 2048
~ 175B parameters

[1] Radford et al.: Language Models are Unsupervised Multitask Learners, 2019.
[2] Brown et al.: Language Models are Few-Shot Learners, 2020.

- see demos at https://transformer.huggingface.co/

In-context learning in GPT-2 and GPT-3

- GPT-2 and GPT-3 ditch the “pre-train and fine-tune” training paradigm of GPT;
- GPT-2 explores unsupervised zero-shot learning, whereas in GPT-3 the authors expand the idea into in-context learning;
- use text input to condition the model on task description and some examples with ground truth.
- Uses zero-shot learning, one-shot learning, few-shot learning (as many examples as they can fit into the context, usually 10-100)
- no gradient updates are performed.

In-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

`Translate English to French:`	*task description*
`cheese =>`	*prompt*

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

`Translate English to French:`	*task description*
`sea otter => loutre de mer`	*example*
`cheese =>`	*prompt*

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

`Translate English to French:`	*task description*
`sea otter => loutre de mer`	*examples*
`peppermint => menthe poivrée`	
`plush girafe => girafe peluche`	
`cheese =>`	*prompt*

Figure source: [Brown et al.: Language Models are Few-Shot Learners, 2020.]

- GPT-3 is still a language model and can be used for text generation

- only 12% of respondents correctly classified this as not written by a human

```
Title:  United Methodists Agree to Historic Split
Subtitle:  Those who oppose gay marriage will form their own denomination
Article:  After two days of intense debate, the United Methodist Church
has agreed to a historic split - one that is expected to end in the
creation of a new denomination, one that will be "theologically and
socially conservative," according to The Washington Post.  The majority of
delegates attending the church's annual General Conference in May voted to
strengthen a ban on the ordination of LGBTQ clergy and to write new rules
that will "discipline" clergy who officiate at same-sex weddings.  But
those who opposed these measures have a new plan:  They say they will form a
separate denomination by 2020, calling their church the Christian Methodist
denomination.
The Post notes that the denomination, which claims 12.5 million members, was
in the early 20th century the "largest Protestant denomination in the U.S.,"
but that it has been shrinking in recent decades.  The new split will be the
second in the church's history.  The first occurred in 1968, when roughly
10 percent of the denomination left to form the Evangelical United Brethren
Church.  The Post notes that the proposed split "comes at a critical time
for the church, which has been losing members for years," which has been
"pushed toward the brink of a schism over the role of LGBTQ people in the
church." Gay marriage is not the only issue that has divided the church.  In
2016, the denomination was split over ordination of transgender clergy, with
the North Pacific regional conference voting to ban them from serving as
clergy, and the South Pacific regional conference voting to allow them.
```

Huge generative language models

- ChatGPT, OpenAI, Nov. 2022
  based on GPT-3.5 with additional training for dialogue
- uses RLHF (reinforcement learning with human feedback)
- demo: https://chat.openai.com/
- huge public impact, possibly disruptive for writing professions, learning, teaching, scientific writing
- GPT-4, 2023: even larger, allows longer context, image input

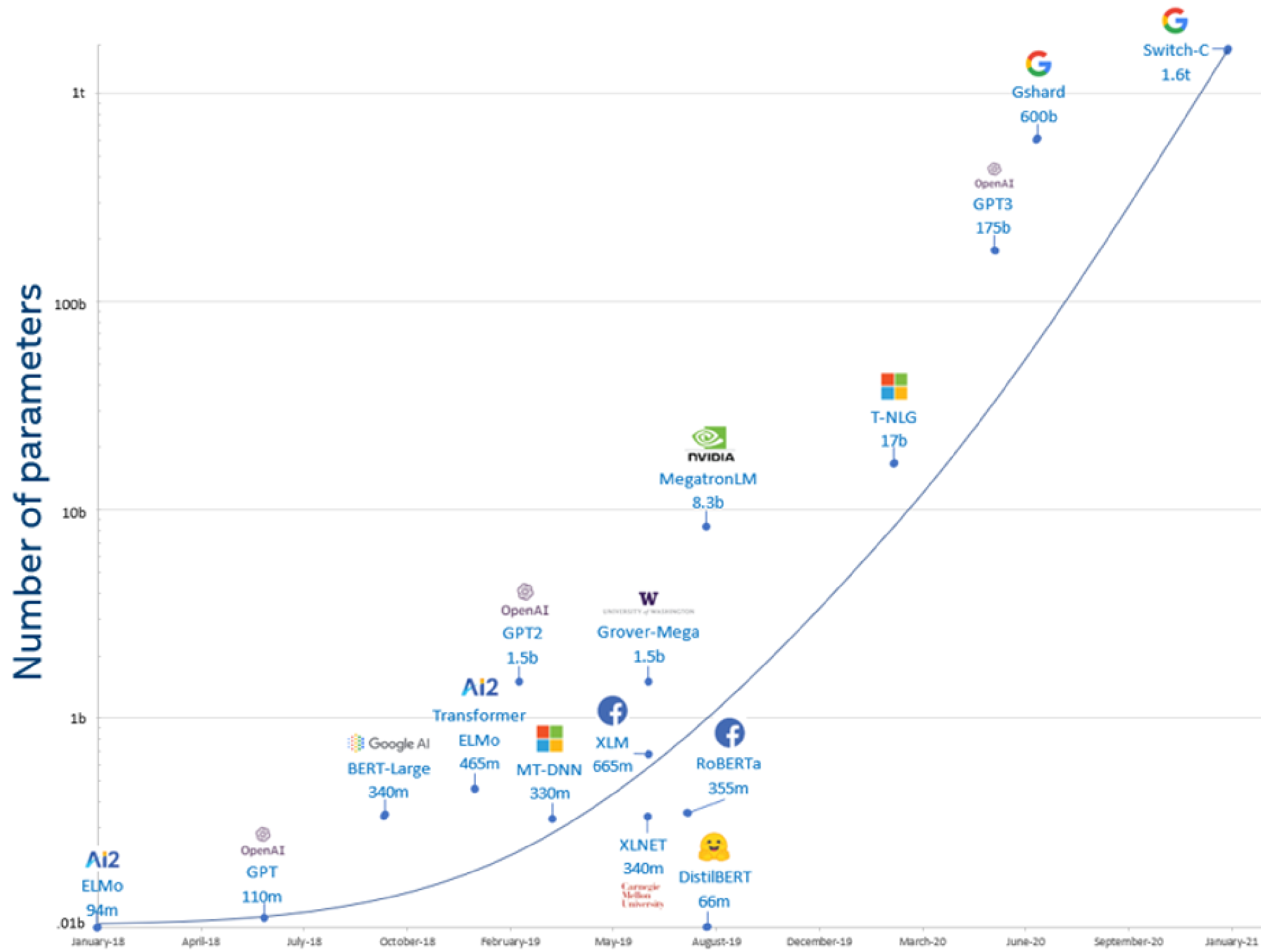


Figure 1: Exponential growth of number of parameters in DL models

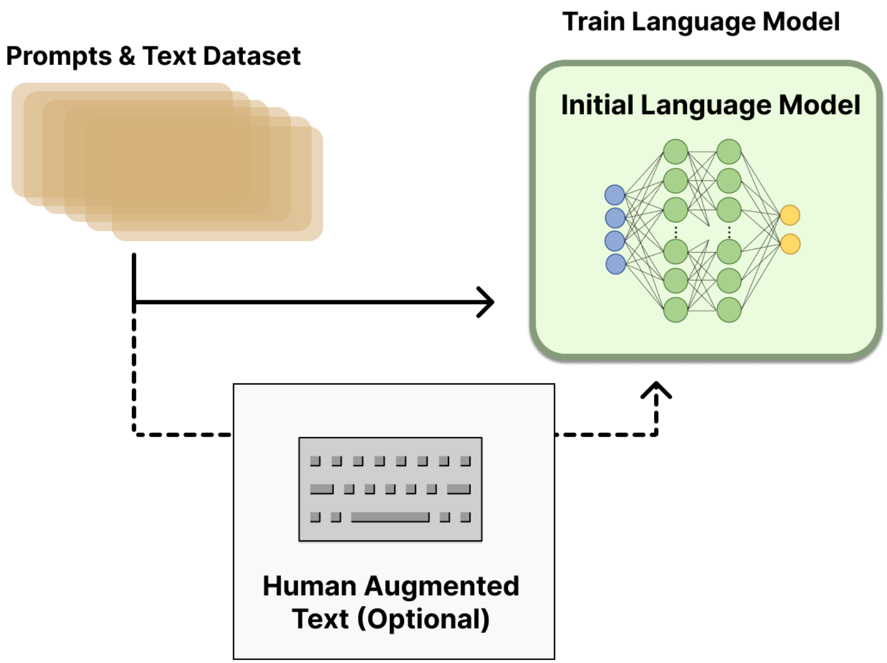
RLHF: idea

- Reinforcement Learning with Human Feedback
- A problem: Human feedback is not present during training
- Idea: Train a separate model on human feedback, this model can generate a reward to be used during training of LLM
- Three stages:
  1. Pretraining a language model (LM),
  2. Gathering data and training a reward model, and
  3. Fine-tuning the LM with reinforcement learning.

Nathan Lambert, Louis Castricato, Leandro von Werra, Alex Havrilla: "Illustrating Reinforcement Learning from Human Feedback (RLHF)", Hugging Face Blog, 2022.

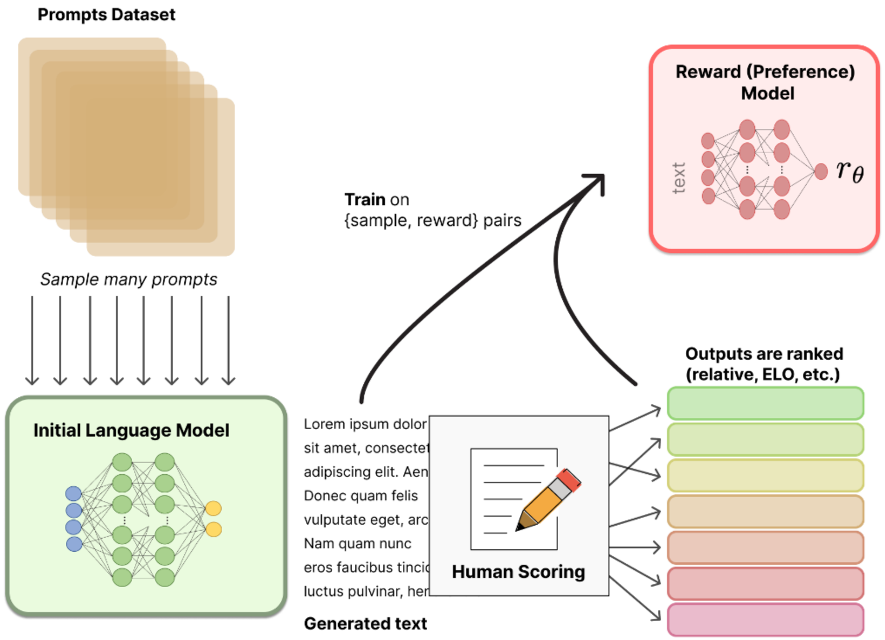
RLHF: the reward model 1/2

- input: a sequence of text, e.g., produced by LM and optionally improved by humans
- output: a scalar reward, representing the human preference of the text (e.g., a rank of the answer)
- the reward model could be an end-to-end LM, or the model ranks outputs, and the ranking is converted to reward



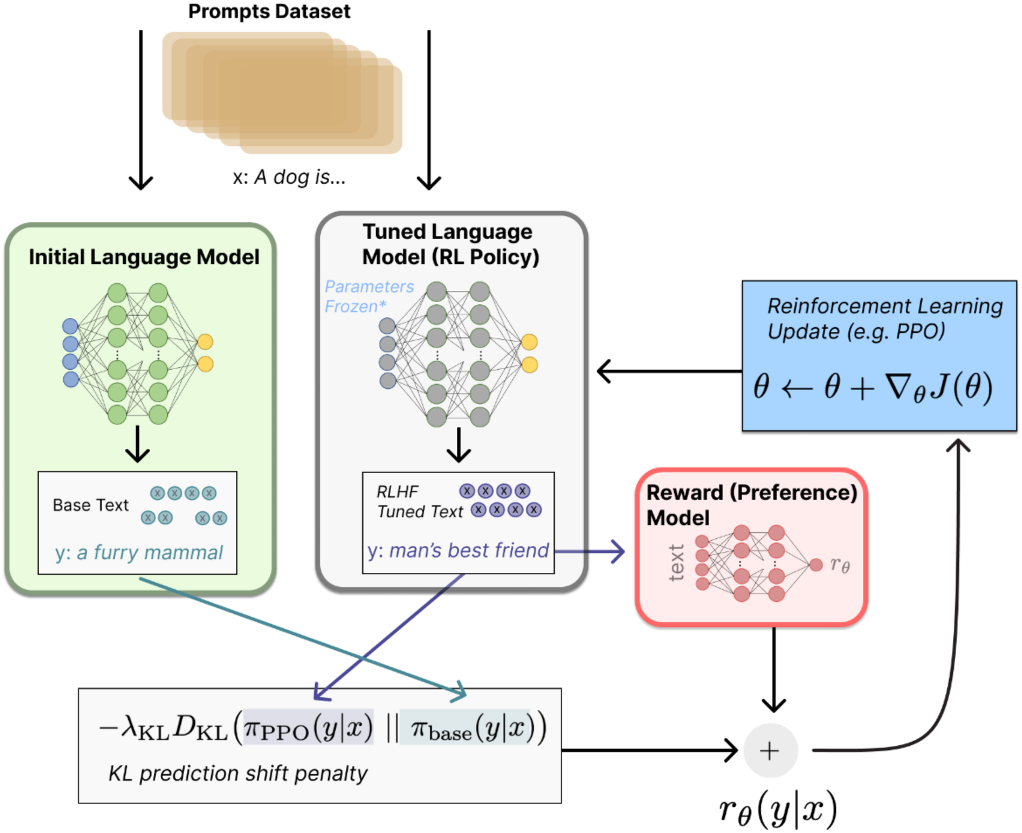
RLHF: the reward model 2/2

- the training dataset are pairs of prompts and (human improved) LM responses, e.g., 50k instances
- humans rank the responses instead of producing the direct reward as this produces better calibrated scores



RLHF: fine-tuning with RL

- RL does not change all parameters, most of parameters are frozen
- the algorithm: Proximal Policy Optimization (PPO)



Attention efficiency

- time and space complexity of self-attention grows quadratically with n (size of input)

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \qquad K, V, Q \in \mathbb{R}^{n \times d}$$

- not suitable for very long sequences like
  - documents
  - character-level language models
  - images (as sequences of pixels);
  - protein sequences.

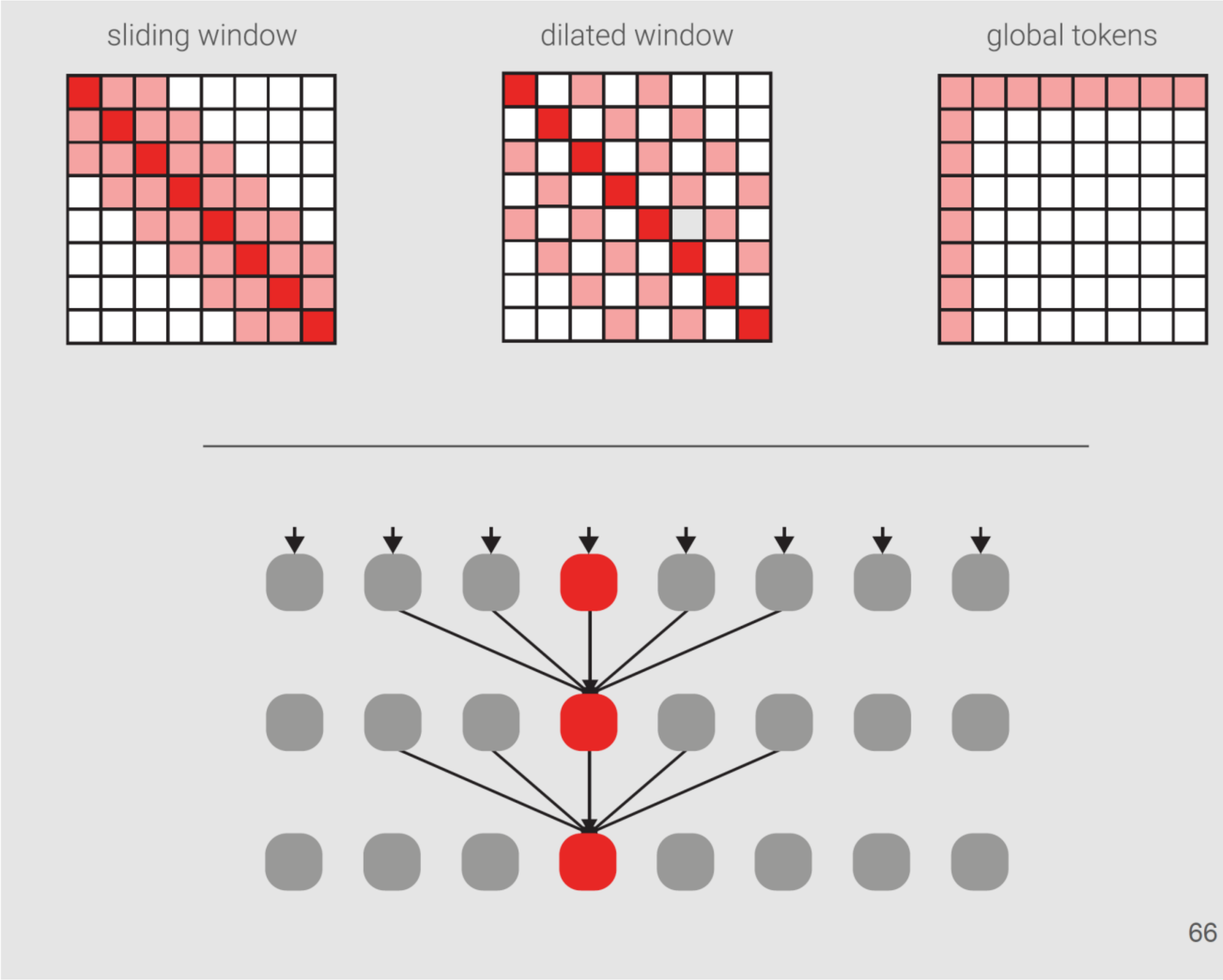
Longformer [1]

- sliding window attention:
each position can attend to 1/2w tokens on each side - **O(w x n)**

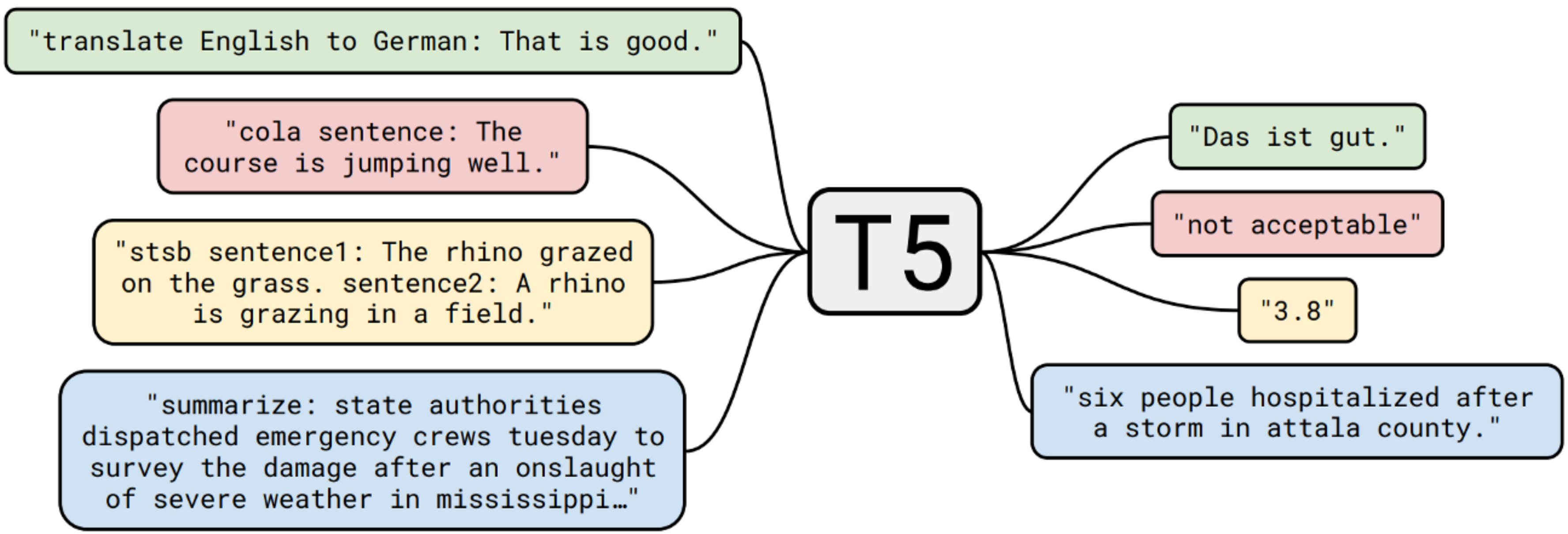
- dilated window attention:
increases the receptive field of the attention layer - **O(w x n)**

- global attention:
k special tokens that aggregate information from whole sequence (e.g. [CLS] as in BERT) - **O(k x n)**

[1] [Beltagy et al.: Longformer: The Long-Document Transformer, 2020.]



T5 (Text-To-Text Transfer Transformer) models



- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y, Li, W. & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, *21*(140), 1-67.
- Ulčar, M. & Robnik-Šikonja, M. (2023) Sequence-to-sequence pretraining for a less-resourced Slovenian language.
- *Frontiers in Artificial Intelligence,* Section on Language and Computation, Volume 6 – 2023, https://doi.org/10.3389/frai.2023.932519

Transformers are everywhere

- music: vocabulary consists of MIDI pitches, pauses, velocity
- object detection (attention to objects)